Building a Causation Annotated Corpus: The Salford Arabic Causal Bank - Proclitics

Jawad Sadek[1], Farid Meziane[2]

[1]National Institute for Health Research Innovation Observatory – Newcastle University
[2]School of Computing Science and Engineering – Salford University

jawad.sadek@newcastle.ac.uk, f.meziane@salford.ac.uk

Abstract

We introduce the Salford Arabic Causal Bank (SACB) corpus, a new corpus dedicated to Arabic *Causal* relations. Causality as a linguistic phenomenon can be expressed using different elements and grammatical expressions. In Arabic language, causal particles – *Purpose Lām*, *Causation Fa'a*, *Causation Ba'a*- are frequently prefixed to words; they play a key role in indicating causality. However, these particles give different meanings according to their position in the text. In fact, these meanings can be interpreted according to the context in which they occur. This ambiguity emphasizes the high demand for a large-scale corpus in which instances of these particles are annotated. In this paper, we present the first stage of building the SACB, which includes a collection of annotated sentences each of which contains an instance of a causal particle. The sentences were carefully examined by two specialist annotators to give an accurate account for each annotated instance. Arabic is a less–resourced language and we hope this corpus would help in building better Information Extraction systems.

Keywords: Arabic Annotated Corpus, Causal Relation, Information Extraction

1. Introduction

Automatic detection of *Causal* relations has gained popularity in the literature within different Natural Language Processing (NLP) applications such as Text Generation, in which causality is exploited to provide explanation and generate knowledge (Kaplan and Berry-Rogghe, 1991). Modern Information Retrieval researchers have focused on developing more efficient search engines by incorporating *Causal* relations into their lexical-based approach (Puente, 2011). Question Answering (QA) is another NLP field to which *Causal* relations is well suited. In particular, it plays a very major part in developing *why*-QA systems (Sadek and Meziane, 2016b; Azmi and Alshenaifi, 2014). Consider for example sentence (1) which contains a *Causal* relation holding between Units 1 and 2. We can return Unit 1 as a candidate answer for the question *"Why was Sarah late?"*

(1) [Because the car broke down,][1] [Sarah was late for school][2].

In Arabic, causality can be expressed using different linguistic elements and expressions. It can be classified into two major categories. The first one is **verbal causality**, which can be captured by the presence of nominal clauses for example, [المفعول لأجله (Accusatives of purpose)-المفعول المطلق (Cognate accusative)] or by causality connectors such as [لذا (therefore), بسبب (because) and من اجل (for)]. The second category is **context-based causality** that can be inferred by the reader using general knowledge without locating any of the previous indicators. This category includes various Arabic stylistic structures that express causality implicitly such as [الاستئناف (resumption), الشرط(condition), الاستثناء (exception)] (Haskour, 1990).

Within the first category, there is a significant group of inseparable particles that are always bound to words. We refer to this group as causal particles, or proclitics for short, and includes: *Purpose Lām* (لام التعليل) – *Causation Fa'a* (فاء السببية) and *Causation Ba'a* (باء السببية).

Arabic authors use these proclitics substantially to indicate causal meaning. In a previous study, we constructed a set of linguistic patterns to detect and extract *Causal* relations expressed in Arabic texts (Sadek and Meziane, 2016a). Several newspaper articles were surveyed in order to design three rule based algorithms that help in recognizing the cases in which the proclitics function as a causative conjunction. Our results reveal that combining the algorithms with the linguistic patterns model has boosted the efficiency by a large margin, improving the overall *recall* measure for Health and Science texts by 29% (out of 195 true positive *Causal* relations, 70 were indicated by proclitics). However, this improvement comes at the cost of *precision* which was reduced by 16% (out of 56 false positive *Causal* relations, 47 attributed to proclitics) i.e. 67% of relations returned by proclitics's algorithms were misclassified. This decline in precision highlights the ambiguity associated with these particles.

The Arabic language, so far, is under-resourced in terms of availability of knowledge base repositories. These resources play an important role in building robust NLP tools and support language technologies' researchers on developing and testing their solutions. Although there are a number of annotated corpora for Arabic, such resources are either 'low-level' (e.g. syntactical or morphological) annotated or they have been labelled with *Causal* relations while annotating other semantic relations. We argue that causation is a complex phenomenon and needs to have annotators to be trained and focus in particular on *Causal* relations.

The syntactical patterns of the Arabic *Causal* relations are rather complex and no general annotated corpus can

provide the diversity of *Causal* relations. So we cannot build on top of any pre annotated corpus but have to create a dedicated corpus of this type of relations. In the current work we introduce the first stage towards building the Salford Arabic Causal Bank (SACB). This stage has been conducted with the goal of collecting and annotating independent sentences where instances of proclitics occurred without regard for other causal indicators.

2. Data Collection

For the purpose of collecting our data, we used the untagged *arabiCorpus*[1] to gather all instances. It is a large corpus consisting of a variety of resources written in Modern Standard Arabic (MSA). The corpus has a Newspapers category containing approximately135 million words of articles published between 1996 and 2010 in different Arabic countries. This category is a good representative for real-world texts as it covers a wide variety of topics.

Searching the *arabiCorpus* for occurrences of words starting with *Lām*, *Fa'a* or *Ba'a*, (henceforth, target word) returns a huge number of matching instances. The issue here is that randomly sampling these instances yields an under-coverage dataset i.e. not every syntactical or semantic form is sufficiently included. This is inherited from the fact that proclitics tend to be highly skewed e.g. the vast majority occurrences of *Fa'a* in Arabic text do not express causation. In which case, most classifiers trained on such dataset would be biased toward major class.

In general, the collected instances must be independent and almost identically distributed. A carefully chosen sample is therefore vital in building a reasonably confident corpus that represents all proclitics' characteristics. To this end, we performed a multistage sampling. We first split the matching instances returned from initial searching (approximately 2.5 million instances) into separate groups according to the length of target words; words of the same length tend to share more linguistic characteristics e.g. grammatical category, morphological pattern. Splitting the data generated five clusters with target word's length of *n*= 2, 3, 4, 5, and over 5 letters; each cluster was then divided into different sub-groups that share one syntactical functionality.

Finally we performed a judgment sampling to avoid data bias. In this phase, the aim is to force the harvested instances to be reasonably balanced between causal and non-causal classes. We requested a native speaker to skim through all clusters and first to randomly select a number of instance that express causation and then to select equivalent number of instances that are non-causal. The number of instances drawn from each cluster was proportionate to the ambiguity of the cluster's population. For example, all instances belonging to clusters of two letters (e.g. لم – في - بث) are classified as non-causal, thus we can be confident that a small size of instances is sufficient to represent these clusters.

3. Annotation Scheme

We used GATE framework (Bontcheva et al., 2013) to support annotation tasks throughout all phases of building our corpus. GATE provides tools for adjudication, integrating multiple annotations set, running various NLP components and supports texts written in Arabic-like script orientation i.e. right-to-left. In addition it permits to create annotation schemas supported by W3C Schema which allows annotation types and features to be pre-specified. In this way, it facilitates the development of Gold Standards. The manual annotation phase was preceded by automatic pre-processing steps. All sentences passed through an NLP components pipeline comprising of the following processes: tokenization, sentence-splitting and POS tagging. We implement the last process using the Stanford POS tagger (Toutanova et al. 2003).

Before an annotation scheme and guidelines can be defined, it is necessary to make clear on what ground we make a judgment on whether the proclitic implies a causal function or not.

3.1 Causal Particles

Causal particles are one of the most complicated and ambiguous particles in Arabic language, as it express many different meaning (Wright et al. 1896). A brief explanation of the particles under consideration in this work is given here.

- **Lām:** It has a multifunctional role and many semantic properties insomuch that some grammarians count more than 30 different purposes of it. For example, (لام الجحود) *Lām of denial* as in "Kalid was not to drink milk" " لم يكن خالد ليشرب الحليب" and (لام الملك) *Lām of possession* when indicating the right of property, e.g. "Ahmad had a large car" " كان لأحمد سيارة كبيرة". However, our concern in this study is *Lām at-'taleel* (لام التعليل) or *Purpose Lām*, which indicates the purpose for which, or the reason why, a thing is done. In this context, the Arab grammarians take *Lām-at-'taleel* to function similarly to (لكي) or (لأن), for example, "he arose to help him" "قام لمعاونته".
- **Fa'a:** It may signal a consequential relationship between two elements or events occurring consecutively, as in "Khalid stood up, then Ahmad" "قام خالد فاحمد". *Fa'a* has also an adversative function, in which it expresses a contrast between two clauses, as in "He invited me, but I turned down his invitation" " دعاني فلم اجب دعوته". In addition, it has a role related to our study in which it contributes to indicating causation between two parts of a sentence, as in "He loved theatre so he excelled in it" "احب المسرح فابدع فيه" (Saeed and Fareh 2006).
- **Ba'a:** It also poses many difficulties. One use of this particle is "الظرفية" to express time and place, for example, "He travelled two days before me" "سافر قبلي بيومين". It can also be used to indicate adhesion "الإلصاق" e.g. "لان الدود يتعلق بالثمار" "because worms stick to the fruit". Another use is to form negation, as in "I don't Know" " لست بعالم". Moreover, it expresses the reason and cause, for example, "كان الاعتداء بقصد السرقة" "The attack committed with intent to steal".

[1] http://arabicorpus.byu.edu/index.php

3.2 Annotation Guidelines

The decision on whether a proclitic serves as a casual indicator may differ according to the way in which it is perceived e.g. syntactic or semantic. In other words, a proclitic which appear to be grammatically a causal particle, the causality may not be contextually perceivable. Since we are dealing with causation from a discourse perspective, we embrace the following principles: *Causal* relation occurs between an event (*the cause*) and a second event (*the effect*) in which the second event is understood as a consequence of the first. When deciding whether there is a *Causal* relation, the annotators were advised to ask whether event B (*effect*) would have occurred if event A (*cause*) had not occurred. If A is a sufficient though not a necessary condition for B to occur, we conclude that A caused B.

A related issue is whether a *cause* or *effect* can be a fact, or whether they have to be an event. In this work, we don't limit *cause* or *effect* to particular types of entities. Thus, an *effect* can be an event, a fact, a method; a *cause* can refer to a reason, motivation, human action, psychological, technological causation etc. We advised annotators to include all the various types. In this context, we label sentences (2) and (3) as two instances holding *Causal* relations indicated by *Ba'a* where the underlined metaphor in sentence (2) represents the *effect*, while the method the *woman* embrace in sentence (3) constitutes the *effect* part of the relation.

(2) نحاول التستر على ضعفنا باخفاء رؤوسنا في التراب، كالنعامة.

"We are trying to cover up our weakness by burying our head in the sand like ostriches."

(3) يتحدث النص عن عجوز تصبر نفسها على الانتظار باسترجاع الذكريات السعيدة من حياتها.

"The text is about an old woman who passes her time waiting by remembering happy moments in her life."

Taking these assumptions into account, the annotators were required to read the entire sentence so that they can make reliable interpretations to the writer's purpose. Then to decide whether the target word indicates a causation based on two facts: both *cause* and *effec*t arguments are securely presented in the sentence where the *effect* has to be explicitly the result of the *cause*; plus each argument constitutes an independent clause i.e. they don't overlap. For example, we classify the particle *Fa'a* in sentence (4) as non-causal. The text does not reveal the fact that made the writer reach his conclusion; and there is no referring expression to any idea mentioned in the previous sentences. As such the reason is only vaguely specified.

(4) لقد قرأت ذات يوم كتابا يقول «كيف تصبح مليونيرا» فلما انتهيت منه ادركت انني لن اصبح مليونيرا.

"I once read a book titled How to Become a Millionaire and when I finished it, I realized that I would never become a millionaire."

It is worth noting that even if the target word indicates causation, the first letter could be a basic unit of the word i.e. it is not a proclitic. The annotators need to be aware of this and should not be tempted to assign a causal status. For example, the target word 'بناء' "at" in sentence (5) starts with *ba'a* that is a part of its original root. The *cause* and *effect* arguments were also annotated if the target word was classified as causal.

(5) التحقت بكلية الحقوق بناء على رغبة امي، فقد ارادت لي ان اصبح محامي مثل والدي.

"I enrolled in the law school at my mother's wish as she wanted me to become a lawyer to follow my father."

Next, the annotators consider a window of five words surrounding the target word and override all POS annotations in this window with new fine-grained ones. This entails assigning different POS tags on sub-word level. The rule-based approach indicates that prefixes and suffixes of surrounding words provide useful hints on proclitics' functionality (Sadek and Meziane, 2016a). All instances annotated according to Stanford POS tag-set, however, we expanded this set so it becomes appropriate to perform fine-grained tagging. For example, we added TIM (ظرف زمان) "adverb of time" - LOC (ظرف مكان) "adverb of place" - PRPY (ضمير متصل) "inseparable pronoun".

The annotators were also required to assign an annotation label referring to the "الوزن الصرفي" "morphological pattern" of the target word. The majority of Arabic words are derived by applying a set of morphological patterns to consonantal roots to which affixes and infixes are added. Morphological patterns are abstractions which can be considered as an indicator of the common concept of the meaning of the word such as *tool* an *event place*/*time* and *instrument*. This classification constitutes a valuable feature in recognizing the role of certain proclitic. For example, a proclitic can be classified as non-causal if the target word belongs to a set of nominal patterns e.g. اسم الفاعل '*present participle*', اسم مكان '*noun of place*'.

3.3 Annotation Process and Adjudication

Two native speakers of Arabic were engaged in the manual annotation process. One annotator (identified as annotator A) was a graduate student in the faculty of Arabic literature. The second annotator (identified as annotator B) was a teaching assistant who has been educated entirely in Arabic. Annotators were trained using the GATE tool on a training set of examples randomly selected from the original dataset. They were asked to identify the function of each proclitic in the training set, and their judgments were compared with the function we had identified in the sentences. We then discussed with each annotator the instances where their judgments differed from ours and clarified the guidelines.

However, it is inevitable that the annotators disagree about the function of some proclitics. In fact, the topic of causation is a matter of debate among experts belonging to this filed (Davidson, 1980; Mackie, 1980). For example, examining the function of the target word "بالتفرج" "looking" in sentence (6), we observed that annotator B assigned causal status to the event " التفرج على لوحاتي " "looking at my drawings", considering the *effect* argument is "keep busy". Annotator A on the other hand conceived the aforementioned event as a request.

In order to create a gold standard set of annotations, we automatically correct all minor mistakes made by annotators using a script written in Groovy language. These corrections are not to interfere or change

annotators' decision, but rather to fix inconsistency e.g. word's length, letter-spacing. We reconciled the differences between annotators by first accepting only instances where both annotators agreed on the binary decision on whether a proclitic indicates causation. Thus we eliminated approximately 300 instances. Then we examined the consensus set for differences in the POS tags. In case there was any disagreement, we included the ones annotated by annotator A as she is an Arabic literature specialist. Table 1 summarizes the main aspects of the final annotated instances: number of instances (N), number of annotated text units (Tokens), number of instances assigned the causal class (causal), number of instances assigned the non-causal class (¬causal). Table 2 illustrates the statistics of instances over the five main clusters. Gate annotation tool format documents in GATE XML style. We converted the documents using another Groovy script so that all annotated instances are encoded in a lightweight XML. Figure 1 provides an excerpt of one instance.

(6) اشغل نفسك بالتفرج على لوحاتي حتى أعد فنجان قهوة وارجع اليك.

"Keep yourself busy looking at my drawings until I make a cup of coffee and come back."

Proclitic	N	Tokens	causal	¬causal
Lām	984	31564	439	545
Fa'a	577	20097	247	330
Ba'a	601	17912	290	311
Total	2162	69573	976	1186

Table 1: Statistics of the dataset

Proclitic	2	3	4	5	+5
Lām	17	61	230	234	442
Fa'a	9	81	111	184	192
Ba'a	22	27	100	114	338

Table 2: Statistics of the dataset based on proclitic's length

4. Related Work

Some research works for Arabic focused on developing annotated corpus with discourse relations. The Arabic Discourse Treebank was generated by (Al-Saif and Markert, 2011) based on the Arabic Penn Treebank. They collected a list of 80 explicit discourse connectives to recognize 18 discourse relations that link adjacent discourse units (DU). The relations are subclasses of four main classes: Temporal, Contingency, Comparison and Expansion. This corpus contains approximately 600 sentences annotated with *Causal* relations under the Contingency class. Another attempt presented by (Keskes et al., 2014) to identify implicit and explicit discourse relations. The authors created an annotated corpus on top of a set of documents extracted from the Discourse Arabic Treebank (Maamouri et al., 2016). The annotation process was performed according to the principles of the Segmented Discourse Representation Theory. They employed the Maximum Entropy model to automatically identify 24 discourse relations holding between adjacent and non-adjacent DUs. The relations were grouped into four top levels classes: Thematic, Temporal, Structural and Causal; of which there are 158 instances annotated with the cause-effect category.

5. Conclusion

There is a lot of uncertainty surrounding the decision about when two events are causally linked. However, the importance and difficulty of extracting causal information suggest that additional efforts are needed in order to reliably create mature language resources. In Arabic, *Causal* relations indicated by causal particles account for a high percentage of the total *Causal* relation in texts. In the current research we created a causation corpus annotated with instances containing words prefixed with certain proclitic along with *cause* and *effect* arguments. In future, we will extend the corpus to include other causal indicators.

```
<Sentence Id="0309" Start="0" End="95">
 <Text>. طلب من اسماعيل ان يأتيه بحجر يكون علامة للناس فذهب اسماعيل يبحث
  عن حجر يؤدي هذا الغرض</Text>
  <Annotations>
   <Annotation Id="11347" Type="Target Word" Start="55" End="58">
     <Features>
       <Length>4</Length>
        <Template>ففعل</Template>
       <Status >causal</Status>
        <String>فذهب</String>
       <Kind>Fa'a</Kind>
     </Features>
   <Annotation Id="11348" Type="Argument" Start="0" End="53">
     <Features>
       <Length>53</Length>
        <String> طلب من اسماعيل ان يأتيه بحجر يكون علامة للناس
</String>
       <Kind>cause</Kind>
     </Features>
   </Annotation>
   <Annotation Id="11349" Type="Argument" Start="58" End="95">
     <Features>
       <Length>38</Length>
        <String> ذهب اسماعيل يبحث عن حجر يؤدي هذا الغرض</String>
       <Kind>effect</Kind>
     </Features>
   </Annotation>
   <Annotation Id="11350" Type="Token" Start="0" End="2">
     <Features>
        <String>طلب</String>
       <Type>arabic</Type>
       <Kind>word</Kind>
       <Length>3</Length>
       <Category>VBD</Category>
     </Features>
    </Annotation>
                  .
                  .
                  .
```

Figure 1: Excerpt from the Salford Arabic Causal Bank.

Acknowledgements

This research work has been funded by Salford University.

6. References

Al-Saif, A., and Markert, K. (2011). Modelling discourse relations for Arabic. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, (EMNLP 2011), Edinburgh.

Azmi, A., and AlShenaifi, N. (2014). Handling "why" questions in Arabic. *In The 5th International Conference on Arabic Language Processing (CITALA'14).*

Bontcheva, K., Cunningham, H., Roberts, I., Roberts, A., Tablan, V., Aswani, N., and Gorrell. G. (2013). Gate Teamware: A Web-based, Collaborative Text Annotation Framework. Language Resources and Evaluation. Volume 47, Issue 4.

Davidson, D. (1980). Causal relations. Essays on actions and events. Oxford University Press. pp. 149-162

Haskour, N. (1990). Al-Sababieh fe tarkeb Al-Jumlah Al-Arabih. Master's thesis, Aleppo University, Aleppo, Syria.

Kaplan, R and Berry-Rogghe, G. (1991). Knowledge-based acquisition of causal relationships in text. Knowledge Acquisition 3, pp 317–337.

Keskes, I., Zitoune, F.B., and Belguith, L. H. (2014). Learning explicit and implicit arabic discourse relations. Journal of King Saud University, computer and Information Sciences vol: 26 (4), PP 398–416.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North America Chapter of the Association for Computational Linguistics on Human Language Technology*, Volume 1. 173-180.

Maamouri, M., Bies, A., Kulick, S. Krouma, S., Gaddeche, Zaghouani,W., 2010b. Arabic Treebank (ATB): Part 3 Version 3.2. Linguistic Data Consortium, Catalog No.: LDC2010T08.

Mackie, J.L. (1980). The cement of the universe: A study of causation. Oxford University Press.

Puente, C., Sobrino, A., and Olivas, J.A. (2011). Retrieving crisp and imperfect causal sentences in texts: From single causal sentences to mechanisms. In *Soft Computing in Humanities and Social Sciences*. pp 175–194.

Sadek, J., and Meziane, F. (2016a). Extracting Arabic causal relations using linguistic patterns. Journal ACM Translations on Asian and Low-Resource Language Information Processing 15, 3, Article 14.

Sadek, J., and Meziane, F. (2016b). A discourse-based approach for Arabic question answering. Journal ACM Translations on Asian and Low-Resource Language Information Processing 16, 2, Article 11.

Saeed, A., and Fareh, F. (2006). Difficulties encountered by bilingual Arab learner in translating Arabic "fa" into English. The International Journal of Bilingual Education and Bilingualism 9, 1, pp 19–32.

Wright, W., and Caspari, C. (1896). A grammar of the Arabic language. Cambridge University Press, Cambridge, England.